An Overview of Some Algorithms in Data Mining

ICTU Student 188
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student188@ictu.edu.vn

***Abstract*—This thesis provides an in-depth overview of data mining, focusing on decision tree and clustering algorithms. It discusses the general concepts and processes involved in data mining, and explores their application in the insurance industry. The author also presents the implementation of these algorithms in a simulation program. The study highlights the importance of data mining in extracting valuable information from large datasets, but acknowledges the limitations due to time and knowledge constraints. The implemented algorithms include the ID3 algorithm for decision trees and K-means and Constrained-K-means for clustering. The author describes the setup of a simulation program for these algorithms, but specific numerical evaluations are not provided.**

***Index Terms*—Data mining, Decision tree, Clustering, K-means, Constrained-K-means, Insurance**

I. Introduction

The thesis introduces the field of data mining, emphasizing its significance in the context of increasing data storage and information needs. Data mining techniques are crucial for extracting valuable information from large datasets. The author discusses the general concepts and processes involved in data mining, including the identification of valuable information, the use of various data types, and the application of different approaches. The study also covers the classification of data mining systems and their applications in real-world scenarios. However, the author acknowledges the limitations of the study due to time and knowledge constraints [1]. Related earlier ICTU undergraduate theses provide context for this English summary [8], [9]. Published articles on the methods and subject of this project provide context for this English summary [10], [11].

II. Methods

The methods section covers the implementation of decision tree and clustering algorithms, including the ID3 algorithm for decision trees and K-means and Constrained-K-means for clustering. The author describes the setup of a simulation program for these algorithms. The ID3 algorithm is used to build decision trees, while K-means and Constrained-K-means are employed for clustering. The author also presents the initialization of the simulation program, including the deletion of existing data and the addition of new data. The implementation of the algorithms is based on the data from the insurance industry, but specific numerical evaluations are not provided [1].

Supervisor: ICTU Teacher 126, Thai Nguyen University of Information and Communication Technology.

III. Results

The thesis presents an overview of data mining and describes ID3 decision trees, K-means clustering, and Constrained K-means. Its conclusion reports a simulation program that applies these algorithms to a problem in the insurance sector. The source appendix includes implementation code for initializing seed sets and cluster centers, which supports the account of a programmed demonstration. It does not provide a measured comparison, an accuracy score, or evidence of adoption by an insurer. The reported result is therefore a simulation and a worked application concept rather than a validated production system [1].

IV. Discussion

The discussion highlights the importance of data mining in extracting valuable information and the potential applications in real-world scenarios. The author also mentions the limitations of the study due to time and knowledge constraints. The study focuses on the implementation of decision tree and clustering algorithms, but the author acknowledges the need for further research and application of these algorithms in practical scenarios [1].

V. Conclusion

The author reports an overview of data-mining concepts and a simulation program for an insurance-related application of ID3, K-means, and Constrained K-means. The conclusion also says that time and expertise limited the depth of the algorithm study. The proposed next step is to improve the application and investigate other methods before practical use. The available source gives no numerical evaluation or real-world deployment evidence. This summary consequently records the program and its scope without presenting the simulation as a proven operational insurance tool [1].

References

[1] ICTU Student 188, "An Overview of Some Algorithms in Data Mining," internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2010.
[2] Nguyễn Đức Cường, Tổng quan về khai phá dữ liệu, Khoa Công Nghệ Thông Tin, Đại học Bách khoa, Tp. Hồ Chí Minh
[3] Semi-Supervised Clustering.Robert Wilkins-School of Engineering and Technology National University
[4] Basu, S., Banerjee, A., & Mooney, R. J. (2002). Semi-supervised clustering by seeding. In Proceedings of 19th International Conference on Machine Learning (ICML-2002), pp. 19–26
[5] Basu, S., Banerjee, A., & Mooney, R. J. (2003). Active semi-supervision for pairwise constrained clustering. Submitted for publication, available at http://www.cs.utexas.edu/˜sugato/

[6] Basu Sugato (2004) Semi-supervised Clustering with Limited Background Knowledge. Proceedings of the Ninth AAAI/SIGART Doctoral Consortium, pp. 979- 980, San Jose, CA, July 2004
[7] Kantardzic Mehmed, 2003 DATA MINING:Concepts, Models, Methods & Algorithms. Copyright © 2003 by Mehmed Kantardzic. pp 79-120.68
[8] ICTU Student 10, "Data Mining: An Overview and Application of Association Rule Mining," internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2007.
[9] ICTU Student 102, "Development of an Intrusion Detection System Based on Data Mining Techniques," internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2009.
[10] D CARVALHO, "A hybrid decision tree/genetic algorithm method for data mining," Information Sciences, vol. 163, no. 1-3, pp. 13-35, 2004. DOI: 10.1016/s0020-0255(03)00414-6.
[11] Rok Rupnik, Matjaz Kukar, Marjan Krisper, "Integrating Data Mining and Decision Support through Data Mining Based Decision Support System," Journal of Computer Information Systems, vol. 47, no. 3, pp. 89-104, 2007. DOI: 10.1080/08874417.2007.11645970.